Biochem. J. (1972) **126**, 683–688
Printed in Great Britain

The Specificity of Dipeptidyl Aminopeptidase I (Cathepsin C) and its Use in Peptide Sequence Studies

By H. LINDLEY
Division of Protein Chemistry, Commonwealth Scientific and Industrial Research Organization, Parkville (Melbourne), Vic. 3052, Australia

(*Received* 8 *October* 1971)

1. The characterization of cathepsin C as an aminodipeptidase has been confirmed. 2. An extra limitation on its specificity, namely that peptides involving proline as the third residue are immune to the enzyme, has been found. 3. A novel technique of using the enzyme for amino acid sequence determination has been shown to be feasible.

McDonald *et al.* (1969) have shown that the enzyme cathepsin C (EC 3.3.4.9) is a dipeptidyl aminopeptidase capable of removing dipeptides sequentially from a peptide chain with an unsubstituted amino terminus. They concluded that 'glycine and proline were preferred *N*-terminal amino acids, whereas substrates possessing an *N*-terminal lysyl or arginyl residue or a penultimate prolyl residue were essentially immune to attack by cathepsin C'. The amino acid sequence of some peptides has been deduced by studying the kinetics of liberation of dipeptides by the enzyme, and the technique has been suggested as a general method of sequence determination for peptides isolated from tryptic digests (McDonald *et al.*, 1969; Callahan *et al.*, 1969). However, the collection of sufficient kinetic results from all but shorter peptide sequences would make this a very complex procedure.

The present paper is concerned with an alternative procedure of sequence determination that uses cathepsin C. In this approach the mixture of dipeptides resulting from a complete digest of the peptide is completely characterized. Another sample of the peptide then has its *N*-terminal amino acid removed, (and incidentally identified) by the Edman procedure and the mixture of dipeptides produced from this cathepsin C is also completely characterized. Deduction of the sequence from this information thus follows the same strategy as is involved in the game of dominoes. In favourable circumstances a unique solution is obtained but ambiguities can arise. The investigation reported shows the practicability of this approach and has described new facts about the specificity of cathepsin C. The Appendix (Rowlands & Lindley, 1972) reports a theoretical investigation of the potential usefulness of the technique with special regard to the degree of ambiguity to be expected.

Materials and Methods

Cathepsin C was prepared from bovine spleen by the method of Metrione *et al.* (1966), and was used essentially under the conditions described by McDonald *et al.* (1969).

S-Carboxymethyl-A- and *S*-carboxymethyl-B-chains of bovine insulin were prepared by the procedure of Thompson & O'Donnell (1966) from a sample of ox insulin supplied by the Commonwealth Serum Laboratories.

The heptapeptide Gly-Phe-Phe-Tyr-Thr-Pro-Lys was isolated from a tryptic digest of *S*-carboxymethyl-B-chain by preparative paper electrophoresis at pH 6.5 (1h, 50V/cm). Despite the apparently complete separation of the peptide from the other B-chain fragments under these conditions, it was found necessary to repeat the electrophoresis on the heptapeptide to obtain an analytically pure sample.

The tetrapeptide Ala-Gly-Pro-Leu was a gift from Dr. F. H. C. Stewart of these laboratories. Edman degradation steps and dansyl *N*-terminal amino acid determinations were done by standard techniques (Blombäck *et al.*, 1966; Gray, 1967; Wood & Wang, 1969).

Amino acid analyses were performed on a Spinco model 120C amino acid analyser, and for this the peptides were hydrolysed *in vacuo* for 18h at 110°C with 6M-HCl in the presence of mercaptoethanol.

Results

Action of cathepsin C on the heptapeptide Gly-Phe-Phe-Tyr-Thr-Pro-Lys

The heptapeptide was incubated for 2h at 40°C with cathepsin C under the conditions used by McDonald *et al.* (1969) and the digest was examined by paper electrophoresis at pH 1.9. This showed that the original heptapeptide had been broken down to three components, which were characterized by amino acid analysis and end-group determination as Gly-Phe, Phe-Tyr and Thr-Pro-Lys. Incubation for short times gave evidence of another component for which amino acid analysis and end-group assay were consistent with the structure Phe-Tyr-Thr-Pro-Lys.

These results are thus in complete accord with the findings of McDonald *et al.* (1969) on the specificity of the enzyme.

Action of cathepsin C on the heptapeptide after Edman degradation

An obvious practical problem arising from the application of the Edman degradation to peptides is that the procedure, in addition to removing the *N*-terminal amino acid also partially blocks the ε-amino group of any lysine residue. This could adversely affect the solubility of the peptide and might interfere with the action of the enzyme.

One solution to this problem in the case of peptides derived from a tryptic digest, would be to remove the *C*-terminal lysine with carboxypeptidase B before the incubation with cathepsin C. However, this may fail with some sequences, and hence it would be useful to have an alternative procedure. Shemyakin *et al.* (1970) claim that phenyl thiocarbonyl groups coupled to ε-amino groups of lysine in peptides can be split off by refluxing with 20% aqueous hydrazine.

It was found that after Edman degradation of the heptapeptide to remove the *N*-terminal glycine two products were formed, which could be separated by paper electrophoresis at pH1.9 (1h, 50V/cm). The mobilities were consistent with one being the hexapeptide and the second minor component being the ε-phenylthiocarbonyl derivative. In confirmation of this view it was found that the second minor component was gradually transformed into the other component by refluxing with 20% hydrazine, the reaction being essentially complete in 45min.

Incubation of the hexapeptide with cathepsin C under conditions identical with those used for the heptapeptide led to the formation of only two products. These were isolated by paper electrophoresis at pH1.9 and shown to be Phe-Phe and Tyr-Thr-Pro-Lys. This result suggests that peptides in which position 3 is occupied by proline may also be immune to cathepsin C, i.e. the enzyme may be unable to split any peptide bond involving proline. Support for this suggestion came from the observation that the tetrapeptide Ala-Gly-Pro-Leu was also immune to the enzyme.

The main results so far reported are shown in Plate 1.

Effect of cathepsin C on the S-carboxymethyl-A-chain of insulin

As a more realistic test of the applicability of the technique to sequence determination we have examined the action of cathepsin C on the *S*-carboxymethyl-A-chain of insulin and the *S*-carboxymethyl-A-chain after removal of the *N*-terminal glycine by the Edman reaction (the des-aminoglycyl *S*-carboxymethyl-A-chain). In contrast with the findings of McDonald *et al.* (1969) on the oxidized A chain it was found that the *S*-carboxymethyl-A-chain was almost completely degraded by cathepsin C as was the des-glycyl *S*-carboxymethyl-A-chain.

Such a complete digestion required overnight incubation at 40°C but it was found that the enzyme could withstand a temperature of 60°C and that at this temperature digestion was complete in 4h and substantially complete even in 2h.

Almost complete separation of the dipeptides in both cases could be achieved by paper electrophoresis at pH3.5 in 2h at 50V/cm, and where further separation was needed this could be achieved by subsequent paper chromatography with butanol–acetic acid–water–pyridine (15:13:12:10, by vol.).

The separated peptides were identified on the basis of *N*-terminal amino acid and complete amino acid analysis, and in this way it was shown that the digest of the A-chain contained the dipeptides Gly-Ile, Ser-Val, Gln-Leu, Leu-Tyr, Val-Glu, Glu-Asn, Tyr-CMCys,* CMCys-Ser, CMCys-Ala and Gln-CMCys.

It also contained free asparagine and the tripeptide Tyr-CMCys-Asn. (The amide groups were not actually assigned in the present work except for the free asparagine, but for present purposes this is an unimportant problem.) The des-glycyl A-chain digest was similarly analysed and found to contain the dipeptides Ile-Val, Ser-Leu, Ala-Ser, Tyr-Gln, Leu-Glu, Glu-Gln, Asn-Tyr, Val-CMCys, CMCys-Asn and CMCys-CMCys. Also the digest contained significant amounts of the tetrapeptide Asn-Tyr-CMCys-Asn.

Thus the enzyme has behaved in accordance with the specificity suggested by McDonald *et al.* (1969) except that there is some indication that a *C*-terminal asparagine residue may be an unfavourable substrate in tri- and tetra-peptide sequences.

Discussion

The above results show that cathepsin C can be used successfully in protein sequence work. For the 21 residue A-chain of insulin the results on the 'odd' and 'even' dipeptides are consistent with only two sequences, namely:

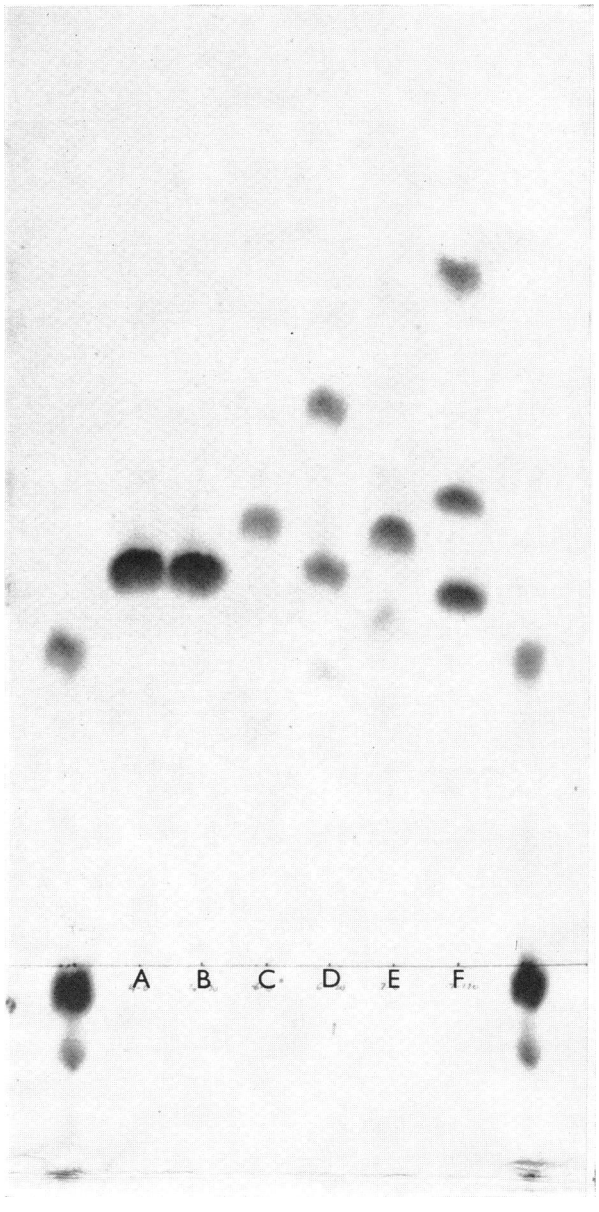
(1) Gly-Ile-Val-Glu-Gln-Cys-Cys-Ala-Ser-Val-Cys-Ser-Leu-Tyr-Gln-Leu-Glu-Asn-Tyr-Cys-Asn

(2) Gly-Ile-Val-Glu-Gln-Leu-Glu-Asn-Tyr-Cys-Cys-Ala-Ser-Val-Cys-Ser-Leu-Tyr-Gln-Cys-Asn

At least two techniques for distinguishing between these alternatives are possible.

(*a*) Studying the time-course of liberation of the dipeptides during digestion by cathepsin C. However, whereas in the techniques suggested by Callahan *et al.* (1969) it is necessary to follow the order of liberation of all the dipeptides, now that we only need distinction between specific sequences only the relative rate of liberation of key peptides need be considered. Thus

* Abbreviation: CMCys (in amino acid sequences), carboxymethylcysteine.



EXPLANATION OF PLATE 1

Action of cathepsin C on proline-containing peptides

Electrophoresis was done at pH 1.9 at 50 V/cm for 1 h with: A, Ala-Gly-Pro-Leu; B, Ala-Gly-Pro-Leu after digestion with cathepsin C for 2 h; C, Phe-Phe-Tyr-Thr-Pro-Lys; D, Phe-Phe-Tyr-Thr-Pro-Lys after digestion with cathepsin C for 2 h; E, Gly-Phe-Phe-Tyr-Thr-Pro-Lys; F, Gly-Phe-Phe-Tyr-Thr-Pro-Lys after digestion with cathepsin C for 2 h. The spots at each edge are from a mixture of marker dyes.

to distinguish between the two alternative sequences given above it would be sufficient to know whether Leu-Glu or CMCys-CMCys appeared first in the digest.

(*b*) A second possibility is to use an enzyme to break down the peptide into smaller fragments that can be investigated. Again, because the problem is one of differentiating between a limited number of known sequences, the choice of enzyme could be straightforward. In the present case chymotrypsin should give either CMCys-Asn or Gln-CMCys-Asn as one of the peptides and the identification of one of these two products should decide which sequence is present. In fact the occurrence of the tetrapeptide Asn-Tyr-CMCys-Asn in the 'even' peptides and the tripeptide Tyr-CMCys-Asn in the 'odd' peptides is sufficient to select the first alternative as the correct one.

The elimination of possible sequences to leave only two alternatives is crucially dependent on being able to distinguish aspartic acid residues from asparagine and glutamic acid from glutamine. This may sometimes pose a problem because the identification needs to be absolute and not merely inference from the net charge on the molecule. Thus the unequivocal identification of Glu-Gln and Glu-Asn is essential to establish the A-chain sequence. This was not done in the present work but could have been achieved by use of the Edman procedure on the significant peptides. However, it is obvious that the techniques used in the present work to separate and identify the dipeptides are too slow and cumbersome to make the sequence procedure attractive in its present form. Nevertheless the practical success achieved, especially when viewed in the light of the theoretical aspects discussed in the Appendix, suggests that if a rapid and reliable method of separating and identifying dipeptides can be developed, the technique will become extremely valuable. An obvious possibility, which it is desirable to pursue further, is the use of mass spectroscopy of suitable derivatives of dipeptides, probably in conjunction with gas chromatography, to effect some preliminary separation. This approach would have the added advantage of coping with the problem of identification of side-chain amides and the whole procedure could ultimately become a standard routine with a computer print-out of the final sequence.

Note added in proof. Since submission of this paper my attention has been drawn to a recent article by McDonald *et al.* (1971), who conclude that 'a prolyl residue constitutes an impasse for dipeptidyl aminopeptidase I if it occurs in a peptide substrate at any position other than at the initial NH_2 (sic) terminus'.

I acknowledge the kindness of Dr. F. H. C. Stewart in synthesizing and making available the tetrapeptide used in this work, and also acknowledge the expert technical assistance of Mr. R. Cranston.

References

Blombäck, B., Blombäck, M., Edman, P. & Hessel, B. (1966) *Biochim. Biophys. Acta* **115**, 371

Callahan, P. X., McDonald, J. K. & Ellis, S. (1969) *Fed. Proc. Fed. Amer. Soc. Exp. Biol.* **28**, 661

Gray, W. R. (1967) *Methods Enzymol.* **11**, 139

McDonald, J. K., Zeitman, B. B., Reilly, T. J. & Ellis, S. (1969) *J. Biol. Chem.* **244**, 2693

McDonald, J. K., Callahan, P. X., Ellis, S. & Smith, R. E. (1971) in *Tissue Proteinases* (Barrett, A. J. & Dingle, J. T., eds.), p. 77, North-Holland, Amsterdam

Metrione, R. M., Neves, A. G. & Fruton, J. S. (1966) *Biochemistry* **5**, 1597

Rowlands, R. J. & Lindley, H. (1972) *Biochem. J.* **126**, 685

Shemyakin, M. M., Ovchinnikov, Yu. A., Vinogradova, E. I., Kiryushkin, A. A., Feigina, M. Yu., Aldanova, N. A., Alakhov, Yu. B., Lipkin, V. M., Miroshnikov, A. I., Rosinov, B. V. & Kazaryan, S. A. (1970) *FEBS Lett.* **7**, 8

Thompson, E. O. P. & O'Donnell, I. J. (1966) *Aust. J. Biol. Sci.* **19**, 1139

Wood, K. R. & Wang, K. T. (1969) *Biochim. Biophys. Acta* **133**, 369

APPENDIX

A Theoretical Investigation into the Potential Usefulness of the Cathepsin C 'Domino' Technique

By R. J. ROWLANDS and H. LINDLEY

Division of Protein Chemistry, Commonwealth Scientific and Industrial Organization, Parkville (Melbourne), Vic. 3052, *Australia*

(*Received* 8 *October* 1971)

In the experimental case of the A-chain of insulin investigated in the main paper (Lindley, 1972), deduction of the sequence from the 'odd' and 'even' sets of dipeptides follows a similar strategy to that used in the game of dominoes and gives only two alternative solutions for this 21-residue peptide. In